PSYCHOVISUAL EVALUATION OF IMAGE SEGMENTATION ALGORITHMS

[1]*Cián W. Shaffrey,* [2]*Ian H. Jermyn and* [3]*Nick G. Kingsbury*

[1]cws23@eng.cam.ac.uk
[1,3]Signal Processing Laboratory, Department of Engineering, University of Cambridge, UK.
[2]Project Ariana (CNRS/INRIA/UNSA), INRIA, Sophia Antipolis, France.

ABSTRACT

Evaluation plays an important rôle in the advancement of any field. In computer vision, unsupervised segmentation algorithms, although of great interest, often suffer from lack of a well-defined goal and/or explicit ground truth data, thus rendering evaluation difficult. This paper presents a novel method for evaluating such algorithms using a database for which ground truth data is not explicitly available. Unlike methods of evaluation that rely on the existence or creation of explicit ground truth data, the proposed evaluation procedure subjects human observers to a *psychovisual* test comparing the results of different segmentation algorithms. The test is designed to answer two main questions: does consensus about a 'best' segmentation exist, and if it does, what do we learn about segmentation schemes? The results confirm that human subjects are consistent in their judgements, thus allowing meaningful evaluation. The relevance of the procedure for the evaluation of CBIR systems is discussed.

1. INTRODUCTION

The task of image segmentation has been much studied in image processing and computer vision. It is often viewed as the first step on the road to statements that concern not the image itself, but the 'scene' of which the image is a representation, this being the goal of machine image understanding. As such, it is important to be able to evaluate the results of image segmentation schemes in a way that does not depend on the opinion of the evaluator.

The output of a segmentation algorithm is a labelled partition of the image domain, or in other words a map from the image domain to a set of labels. The status of the set of labels divides segmentation algorithms into two categories. The label associated with each pixel may imply a statement about a quantity in the scene of which the image is a representation (this is equivalent to having a well-defined 'semantic space', as defined and discussed in [1]), or it may not. Note that this distinction is different from that between supervised and unsupervised segmentation schemes. A scheme in which models for classes in the image are trained beforehand, or in which regions are selected in the data at hand and used to train classes, without assigning a semantic value to those classes, fall into the second category. On the other hand, a clustering algorithm that assigns values to its parameters based only on the data at hand with no interference from a human operative could fall into the first class if it is asserted that the aim is to segment the image into such-and-such semantically pre-defined possibilities.

Although there may be problems in actually obtaining the 'ground truth' information for the first category of schemes, if ground truth can be obtained then in principle there is no problem with their evaluation. It is possible to compare the output of the segmentation scheme with 'measurements' (in a generalised sense) of the scene, and to determine whether or not the statements implied by the label assignments are true. A metric must be chosen with which to compare the output of the scheme to the ground truth, but this is a technical problem rather than a conceptual one. Previous work has taken this approach to image collections for which well-defined semantics exist, but for which ground truth is hard to obtain. Chalana and Kim [2] and Yang *et al.* [3] use multiple expert observers to agree on ground truth in the context of medical imagery, while Hoover *et al.* [4] do so in computer vision.

The same is not true of the second category of schemes however. In the absence of well-defined semantics for the segmentation, one does not have a reference point to which to compare the output of the scheme. The absence of well-defined semantics may simply be the result of failing to define some. In many cases however, it is a result of the complexity of the images involved. For many classes of images, the semantics may seem unbounded and the task of defining them too difficult. In this situation, which is not at all rare, is it then possible to perform an evaluation of segmentation schemes at all? The performance of just such an evaluation is the subject of this paper.

1.1. Evaluation Methodologies

Despite this analysis, the idea persists, implicit in segmentation schemes that have no well-defined semantics, that some ways of partitioning the image domain are better than others

This work was supported by EU project MOUMIR (HP108), www.moumir.org

for most choices of statement one may wish to make about the scene. In the absence of explicit semantics, the only alternative is to turn to human subjects, who will introduce implicit semantics through their understanding of the images. The critical question in any such endeavour is whether a consensus about the evaluation emerges. If it does not, then it is hard to claim that the notion of a segmentation in the absence of explicit semantics is amenable to evaluation at all. If a consensus does emerge, then the nature of that consensus is in itself the most interesting consequence of the evaluation. It can then be used to compare different segmentation schemes.

One approach is to ask human subjects to segment the images by hand. If a reasonable consensus emerges, the hand segmentations can be treated as ground truth, and compared to the outputs of segmentation schemes. Martin *et al.* [5] take this approach. They have human subjects hand segment images from the Corel database, and then analyse the resulting segmentations. They indeed find a degree of consistency across the image data set that they use.

An alternative approach is to allow human subjects to evaluate directly the output of segmentation schemes using psychovisual tests. This has two advantages. It allows the minimum of instruction to be given to the subjects, thus permitting them to use whatever semantics seem most natural to decide which of the segmentations seems most meaningful to them. Second, it can be used even when the generation of hand segmentations is difficult due to the semantic complexity of the images, as is the case with the testbed of images used in this paper.

Within this approach, it is at first natural to think of asking the subjects to assign an absolute value to the segmentations from the different schemes. One could then compare these values. Unfortunately the meaning of these values would be very hard to define, and might differ from image to image throughout a sequence of test images. The results would thus be very hard to analyse. The approach taken in this paper is a little different. Human subjects, rather than giving absolute values to segmentations, are instead asked to choose between the segmentations resulting from different schemes. In addition to the binary variable indicating their choice, the time taken for them to reach their decision is recorded. The intention is to discover an underlying consensus about the *difference* in meaningfulness of the different segmentations. The subjects express a preference from two schemes at a time on a number of images. Analysis of the results, described in section 4, then leads to a pairwise ranking of the schemes. It is not necessary that such a pairwise ranking be consistent with any single total order on the schemes: cycles may exist in the pairwise ranking rendering this impossible. It is a first check that consensus exists that such a total ordering is in fact possible.

The impetus for this work arose from the area of content-based image retrieval (CBIR), and in particular the evaluation of retrieval systems. (See [1] for a discussion of evaluation methodologies for retrieval systems.) CBIR has become increasingly important as a research area in recent years, due to the explosive growth both of image archives in many fields and of the Web. It is also an interesting area theoretically, bringing all the important questions of machine vision into focus and providing a natural and potentially objective testbed for image understanding systems. CBIR attempts to replace current retrieval based on manual textual annotation of images with automatic processing in which indices are generated for each image and then used for retrieval. For this to work well, the indices must capture the relevant semantics of the images in the database, and it is here that the link with segmentation methods arises. It is clear that for many, maybe even most, queries to such systems, a *sine qua non* of successful retrieval will be a description, at some level, of the 'principle objects' in the images in the database. The phrase 'principle objects' is of course not well-defined and may even vary from query to query. If such segmentations do exist however, then the ability to produce them automatically will be a good first indicator of success in a retrieval setting.

The paper is structured as follows. In section 2 we describe the basis on which the segmentation schemes and the images used in the evaluation were selected. A short description of each scheme is also presented. In section 3 we describe in detail the evaluation procedure. Section 4 presents the results obtained from the evaluation procedure and we discuss these and conclude in section 5.

2. IMAGE SEGMENTATION SCHEMES AND DATASET

Any scheme for evaluating segmentation methods must choose a test-bed of images with which to work. Conclusions drawn using this testbed will not *a priori* generalise to other types of images. Such a conclusion can only be reached after extensive experimentation with many types of images. Nevertheless, the analysis of the results from any one testbed may suggest avenues of exploration both in research on segmentation algorithms and on evaluation methodologies.

The images used for evaluation in this paper are scanned images of fine art paintings from the Bridgeman Art Library (BAL) collection.[1] BAL is a commercial art library supplying electronic and hard copy images to magazines, newspapers, designers and others. The images are realistic in intent, but in many cases the colours and forms do not correspond to 'photographic realism'. The semantic content of the images is complex and varied. Some are landscape scenes, while others are portraits. Very often there is no dominating object in the foreground on which attention can be fixed and

[1] All images in this paper are used with the permission of the Bridgeman Art Library, `www.bridgeman.co.uk`.

that can be used to label the image as a 'picture of X'. It is clear that the semantics of these images are more or less unbounded, and that the idea of defining 'ground truth' is highly optimistic. The collection thus falls squarely into the first category mentioned in the introduction. In addition, it is hard to generate hand segmentations of the BAL images. The number of possibilities is extremely large, and our attempts to generate such segmentations were a failure. Subjects did not know what to do.

Two subsets of 10 and 50 images respectively were selected at random from an initial set of 3000 images from the BAL collection. Although the segmentation schemes are unsupervised, some minimal setting of parameters is required. This is to avoid drastically over- or under-segmenting the images, thus adding further difficulty to the evaluation problem. Thus, the first subset of 10 training images was used to fine-tune the parameters of the schemes. The fine-tuning was done by the authors of each scheme. Each image was accompanied by rough guidelines indicating the number of regions expected, and some idea as to the identity of the principle regions in each image. The number of regions ranged from 2–10. After the fine-tuning, the parameters were left untouched while the schemes segmented the remaining 50 test images. These 50 images and the resulting segmentations were then used in the evaluation.

Six schemes were made available for evaluation, of which five came from the members of the MOUMIR project and one from outside (Blobworld). Details of the features used, the models and the algorithms can be found in the cited papers.

- *Multiscale Image Segmentation (MIS)*: This scheme is outlined in [6]. It generates classes using a robust mean shift procedure, which operates on a 7 - dimensional joint spatio-feature space, containing 3 colour, 2 texture and 2 spatial feature components. The subsequent classification procedure consists of a Bayesian multiscale process which models the inherent uncertainty in the joint specification of class via a Markov Random Field model.

- *Blobworld*: The Blobworld scheme aims to transform images into a small set of regions which are coherent in colour and texture [7]. This is achieved by clustering pixels in a joint colour-texture-position feature space using the EM-algorithm.[2].

- *Iterated Conditional Modes* (ICM): The likelihood for this model uses an i.i.d. Gaussian model of pixel intensities, combined with a Potts-like prior. ICM is used to maximise the posterior probability, an initial configuration being created using k-means. A full description can be found in [8].

[2]It was not feasible to alter any of the internal parameters of Blobworld.

- *Learning Vector Quantization* (LVQ): The feature vectors used in the LVQ clustering algorithm consisted of the RGB colour values and the coordinates of each pixel. LVQ is a self-organizing neural network with a competitive learning law. The model is described in [9].

- *Double Markov Random Field* (DMRF): The double Markov random field assumes Gaussian MRF models for classes within the image, and that class labels follow a Potts model [10]. In this approach, the posterior distribution of class labels and all model parameters given observed intensities are simulated using a Markov chain Monte Carlo approach. The segmentation is taken to be the marginal posterior mode, where each pixel is classified to be that class that was sampled most often in the simulation.

- *Complex Wavelets and Hidden Markov Trees* (CHMT): Segmentation using Complex Wavelets and Markov Trees is initialised using the mean shift procedure to generate classes. Then texture and colour models, based on hidden Markov trees of complex wavelet [11] and scaling function coefficients respectively, are trained. A segmentation is found by using maximum likelihood classification of the coefficients given the models [12].

An idea of the variation in the segmentations produced by the schemes that we evaluate can be obtained from figure 1(b), which shows the output of six different segmentation schemes given the image in figure 1(a).

3. EVALUATION METHOD

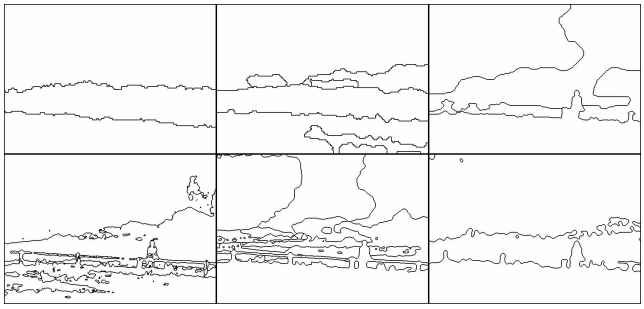
The psychovisual test consisted of a series ('evaluation set') of 'trials', each of which asked the subject ('subject') to choose between the segmentations of a single image by two different schemes. As stated, six schemes were compared to each other, thus giving 15 pairwise comparisons. The number of subjects was 14.

3.1. Evaluation Set

Previous work in the area of psychovisual testing [13] suggests that a 30-minute time limit should be placed on the length of the test. This is to guard against subject fatigue, which could influence the results in an unpredictable manner. In preliminary investigations of the testing process, we found that each trial took about 10 seconds. An evaluation set therefore consisted of 150 trials. The existence of 15 pairwise comparisons meant that each pair of schemes was compared over 10 trials.

To assign 10 images to each pair of schemes, we sampled without replacement from the 50 images in our test set, beginning again when all images had been used. The result-

(a)



(b)

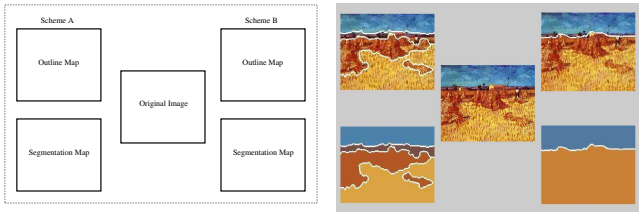
Figure 1: *1(b) shows the variation in the segmentations of the image in 1(a) resulting from the six schemes evaluated (Image © Bridgeman Art Library).*

ing 150 trials were then randomly sequenced, and the two schemes in each trial assigned randomly to left or right for display purposes (see section 3.2). Once done, this same evaluation set was used for each subject.

3.2. Trial

In each trial, five images were used. The original image was displayed centre screen, with, on either side of it, the segmentation results from the two schemes being compared in that trial. Each scheme's segmentation was presented to the subject using two different representations. We term these images the 'outline map' and the 'segmentation map'. The outline map shows the region boundaries superimposed on the original image. The segmentation map shows the regions themselves by colouring a region with the mean colour of the regions corresponding to its class. Figure 2 illustrates the layout of a trial.

The reason for using two representations of a single segmentation result is the following. The outline map is designed to display the coincidence (or not) of region boundaries with features in the image, and the nature of the interior of each region. The segmentation map is designed to show the global



(a) (b)

Figure 2: *2(a) is a schematic of the on-screen layout of a trial, while 2(b) contains a screen shot (Image © Bridgeman Art Library).*

(a) This screen shot shows the first stage of a trial, in which the original image is displayed to the subject until they are familiar with it. To proceed to the trial, the subject clicks on the original image.

(b) This screen shot shows the second stage of a trial, in which the segmentations from the two schemes are represented by their outline and segmentation maps. This stage of the trial is timed, resulting in the soft score described in the text.

Figure 3: *This figure contains the two stages of a trial (Image © Bridgeman Art Library).*

layout of the regions and to show which regions are classified as belonging to the same class. It also presents a greatly simplified version of the original image.

The images in each trial were displayed in two stages. First, the original image was displayed on its own. The subject had an indefinite amount of time to look at this image and become familiar with it. When ready, the subject could click on the original image, at which point the other four, segmented images were displayed. When the subject had made a decision, the next original image was displayed, thus commencing the next trial. Figure 3 shows these two stages of a trial.

During each trial, two types of measurements were made of the subjects' response to the proposed segmentations: a 'soft score' and a 'hard score'. The hard score (± 1) indicates which scheme the subject preferred, and the soft score multiplies this by the speed of response for the second stage of the trial. The time taken looking at the original image be-

fore the segmented images were displayed was recorded, but is not used in this paper. The idea of displaying the original image first, and then of waiting until the subject is ready to proceed, was to reduce as far as possible the effects of image complexity on this soft score. In practice, for analysis, we use the reciprocal of the time, termed the 'speed'. The speed is signed, the sign of the speed indicating which of the two schemes the subject chose. If the subject was 'undecided', the speed was set to zero. The measurement of speed is intended to indicate how close in meaningfulness the segmentations from the two schemes are to each other (although not necessarily close geometrically), a longer decision indicating that the two segmentations were closer. This type of 'speed of response' measurement has proven to be an effective measure in other psychovisual tests relating to detecting image compression artifacts [13]. The hard score consists simply of the sign of the speed, therefore indicating which scheme the subject chose, but not how quickly. These two measurements form the basis of the results analysis in section 4.

3.3. Instructions

Prior to performing the psychovisual tests, all subjects were issued with a set of instructions. The instructions were designed to be minimal, in the sense of influencing the semantics the subject would use to understand the image as little as possible. The instructions read: *The pair of images to the left of the original image illustrates one way of splitting the original image into its most important pieces, while the pair of images to the right of the original image illustrates a second way. Decide which of the ways, left or right, of splitting the image into its most important pieces makes most sense to you*. Once the subject comes to a decision, using a mouse they click on either of the outline or segmentation maps of the chosen scheme. Subjects may also be *undecided* in their choice, in which case they click on the original image.

In order to allow subjects to familiarise themselves with the test, each subject underwent a short 'familiarisation session' prior to performing the test proper. The session consisted of 10 trials. The images used in the familiarisation session were not used again.

4. RESULTS AND DISCUSSION

The first goal of the analysis is to determine whether or not there is coherence in the results of the psychovisual tests. In particular, are they consistent with a total order on the schemes? To obtain such answers, the measurements obtained from the 14 different subjects who took part in the evaluation need to be combined in some way. Recall from section 3, that during each trial two measurements were made: 'hard' and 'soft'. The soft measurement is the (signed) speed of the subject on that trial, while the hard score, the ternary choice the subject made, is equal to the sign of the speed: $+1$ to the 'winning' scheme, -1 to the 'losing' scheme, and 0 to both if the decision was 'undecided'.

Hard Scores

	ICM	CHMT	DMRF	LVQ	Blob.	MIS
ICM	–	0.04	0.13	0.09	0.09	0.07
CHMT	-0.04	–	0.1	0.14	0.19	0.11
DMRF	-0.13	-0.1	–	0.14	0.27	0.04
LVQ	-0.09	-0.14	-0.14	–	0.11	0.34
Blob.	-0.09	-0.19	-0.27	-0.11	–	0.11
MIS	-0.07	-0.11	-0.04	-0.34	-0.11	–

Table 1: *The table shows the hard scores produced by the procedure described in the text. Each row of the table contains the scores of that scheme against the others. For example, the ICM scheme scored* 0.13 *against DMRF,* 0.09 *against LVQ etc. A positive score in a row is good for that scheme.*

In order to compare the results from different subjects, some model has to be given for the variation among subjects. In principle this variation could be extremely complicated, in which case there is little hope of discovering a consensus. We make perhaps the simplest assumptions possible: that for each trial, or in other words for each image and each pair of schemes, there is an 'intrinsic' score d that determines each subject's mean speed linearly. In other words, each subject possesses a 'speed coefficient' α that multiplies the intrinsic score to give the mean speed. We further assume that the observed speeds are distributed normally about this mean with a subject-dependent variance, $\sigma^2 = 1/\lambda$. By assuming ignorance priors on α, λ and d, marginalising over α and λ, and then taking a MAP estimate, we arrive at the simple recipe of first normalising the vector of speeds of each subject over all trials using the Euclidean norm and then averaging over subjects, giving an intrinsic score for each trial. We average the hard scores over subjects also.

In order to create a pairwise ranking of the schemes, we average the intrinsic scores over the trials for each pair of schemes, giving a hard and a soft score for each pair of schemes. One can view this as 'averaging over the dataset'. These scores give us two pairwise rankings for the set of schemes, which may or may not be consistent with each other. The results of this procedure for the hard scores are shown in table 1, while those for the soft scores are shown in table 2.

Note that, if s is an entry in table 1 for schemes (A,B), then $(1+s)/2$ is the frequency with which scheme A was picked. Some of the frequencies are not far from 0.5 at first sight, but it must be borne in mind that they are averages over 140 samples. The probability of a frequency of 0.45 being produced by a binomial distribution with $p = 0.5$ is small (~ 0.03). Another way of saying the same thing is that the standard

Soft scores

	ICM	CHMT	DMRF	LVQ	Blob.	MIS
ICM	–	5.3	11.3	9.2	11	9
CHMT	-5.3	–	5.8	14.8	15.6	20.6
DMRF	-11.3	-5.8	–	6.9	31.7	0.5
LVQ	-9.2	-14.8	-6.9	–	6.9	35.1
Blob.	-11	-15.6	-31.7	-6.9	–	17.6
MIS	-9	-20.6	-0.5	-35.1	-17.6	–

Table 2: *The table shows the soft scores produced by the procedure described in the text, multiplied by* 1000*. Each row of the table contains the scores of that scheme against the others. For example, the ICM scheme scored* 11.3 *against DMRF,* 9 *against MIS etc. A positive score in a row is good for that scheme.*

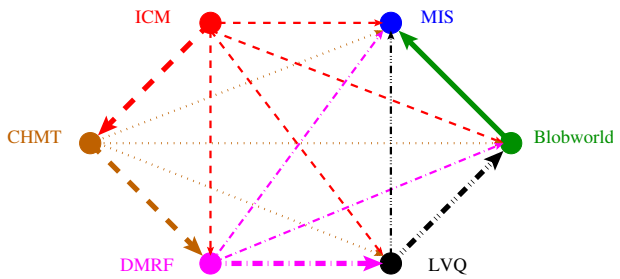


Figure 4: *A graph illustrating the pairwise ordering arising from both the hard and soft scores, and the resultant total ordering. There is an arrow from vertex A to vertex B if the scheme associated with vertex A performed better than that associated with vertex B. The arrows from a given vertex are coloured and dashed in the same way, and the total ordering arrows are shown thickened.*

deviation of the MAP estimate (0.45) of p for a binomial distribution when 140 observations give a frequency of 0.45, is 3.4×10^{-3}. The randomization procedures used to construct the experiment remove bias as far as possible, and we are left with a significant effect.

Given the results, the following natural questions arise: do the two (hard and soft) pairwise rankings lead to consistent total orderings? Are these total orderings the same for the hard and soft scores? In both cases the answer is 'Yes'. Figure 4 shows a graph in which an arrow from vertex A to vertex B indicates that the scheme associated with vertex A performed better than that associated with vertex B. The diagram illustrates how unlikely it is *a priori* that such a pairwise assignment would lead to a total order. For six schemes, there are approximately 45 possible pairwise orderings for each possible total ordering. The order resulting from this analysis is: ICM $>$ CHMT $>$ DMRF $>$ LVQ $>$ Blobworld $>$ MIS.

The results are intriguing, because ICM has the simplest feature set of any of the algorithms tested, eschewing even the use of colour. CHMT on the other hand has one of the most complex, training hidden Markov tree models of wavelets for texture, and using scaling coefficients for colour. One explanation might be that the coincidence of region boundaries with semantically significant boundaries in the image is an important factor determining subjects' responses. The use of greyscale intensity differences in the ICM model renders it sensitive to sharp boundaries that models that include texture may miss, thus helping to compensate for the lack of sophistication in its features.

Perhaps the closest direct comparison is between ICM and LVQ. Both use simple features (ICM uses pixel intensity, LVQ pixel colour), but they differ in the way they take into account region geometry. LVQ clusters pixels using their coordinates as another feature, while ICM applies a Potts-like prior favouring consistent regions. It would appear that the latter is more successful.

Connected to this difference is the possibility that the number of regions (as opposed to classes) plays an important rôle. At least as limiting cases (one region, or a very large number), it is clear that this is a significant factor. LVQ tended to produce a large number of regions, whereas ICM produced a reasonable number (say ~ 5). This is not a consistent interpretation of the results however, as Blobworld also produces a reasonable number of regions. However, Blobworld's use of an *a priori* structure for these regions may have hampered its performance on the images in the BAL database. Blobworld was originally applied to natural images often possessing a single dominant object in the centre foreground.

The results give an indication that a consensus about 'fundamental' image segmentations exists. Further experiments and analysis are necessary to understand the nature of this segmentation.

5. CONCLUSION AND DISCUSSION

The evaluation of segmentation schemes in the absence of well-defined semantics is a thorny problem. Without explicit knowledge of what one would like the output of the algorithm to be, it is hard to say whether one scheme is better than another. The most important question for such schemes is whether there exists any kind of consensus arising perhaps from a hidden and somewhat universal set of semantics. The only way to access such a set if it exists, is to perform tests with human subjects. One way would be to allow subjects to create hand-segmentations, and look for consistency amongst the results, subsequently using the results as 'ground truth'. In this paper we have taken another approach, asking subjects to compare directly the output of segmentation algorithms and judge which of a pair of segmentations is more meaningful to them. This has the advantages that

it allows the minimum of instruction to be given to the subjects, and it can be used even when the generation of hand segmentations is difficult due to the semantic complexity of the images. The images used for evaluation in this paper are scanned images of fine art paintings, which are very complex semantically.

Based as they are on pairwise comparisons of segmentation schemes, it is remarkable that the results we obtain are consistent with a total ordering on the six schemes tested, and that this ordering is essentially unaltered by various means of analysing the data. Certainly the results are very far from chance levels. This consistency suggests that human subjects do perceive images as broken up into regions in a consistent way, and that further study might enlighten us as to what criteria are at play in this process. The results of our study are inconclusive about the reasons for the ordering we obtain. The two most successful schemes use very different features and models. Further testing and analysis is necessary to determine if there are commonalities linking the successful segmentations.

Central to this evaluation procedure is the hypothesis that an image segmentation scheme which closely mimics human interpretation of semantic content is in a better position to attempt retrieval of that content within a CBIR setting. Thus, work is currently under way integrating the CHMT scheme into a CBIR framework, results of which will be published at a later date.

Acknowledgements

The authors wish to express their thanks to the Bridgeman Art Library for their permission to use the images in the paper. They also wish to thank to Dr. Erinija Pranckeviciene from the Artificial Intelligence and Information Analysis Laboratory of the Aristotle University of Thessaloniki, Greece, and Dr. Simon Wilson from the Statistics Department of Trinity College Dublin, Ireland, who together provided three of the segmentation schemes used in the evaluation process. In addition, the authors would like to thank the many members of the Signal Processing Laboratory of the Cambridge University Engineering Department and Project Ariana (CNRS/INRIA/UNSA) at INRIA Sophia-Antipolis for agreeing to act as subjects. This work was supported by European Union project MOUMIR – `www.moumir.org`

6. REFERENCES

[1] I H Jermyn, C W Shaffrey, and N G Kingsbury, "Evaluation methodologies for image retrieval systems," in *Proc. Advanced Concepts for Intelligent Vision Systems*, Ghent, Belgium, September 2002.

[2] V Chalana and Y Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Trans. on Medical Imaging*, October 1997.

[3] L Yan, F Albregten, T Lønnestad, and P Gtøttum, "A supervised approach to the evaluation of image segmentation methods," in *Proc. Int'l. Conf. on Computer Analysis of Images and Patterns*, Prague, Czech Republic, September 1995.

[4] A Hoover, G Jean-Baptiste, X Jiang, P Flynn, H Bunke, D Goldgof, K Bowyer, D Eggert, A Fitzgibbon, and R Fisher, "An experimental comparison of range image segmentation algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, July 1996.

[5] D Martin, C Fowlkes, D Tal, and J Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int'l. Conf. Computer Vision*, Vancouver, Canada, July 2001.

[6] A Kam and W Fitzgerald, "A general method for unsupervised segmentation of images using a multiscale approach," in *Proc. European Conf. on Computer Vision*, Dublin, Ireland, June 2000.

[7] S Belongie, C Carson, H Greenspan, and J Malik, "Color and texture-based image segmentation using the EM algorithm and its application to content-based image retrieval," in *Proc. IEEE Int'l. Conf. on Computer Vision*, Bombay, India, 1998.

[8] T Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. on Signal Processing*, April 1992.

[9] C Kotropoulos, E Auge, and I Pitas, "Two-layer learning vector quantizer for color image segmentation," *Signal Processing IV: Theories and applications*, 1992.

[10] D Melas and S Wilson, "Double markov random fields and bayesian image segmentation," *IEEE Trans. on Signal Processing*, February 2002.

[11] N G Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Journal of Applied and Computational Harmonic Analysis*, May 2001.

[12] C W Shaffrey, N G Kingsbury, and I H Jermyn, "Unsupervised image segmentation via markov trees and complex wavelets," in *Proc. IEEE Int'l. Conf. On Image Processing*, Rochester, NY, September 2002.

[13] S Karunasekera and N Kingsbury, "A distortion measure for blocking artifacts in images based on human visual sensitivity," *IEEE Trans. on Image Processing*, June 1995.